Sparse Autoencoders Find Causal, Lineage-Specific Context Features in Chromatin Foundation Models

Nicole Ching [1] **Ayushi Mehrotra** [1]

Abstract

Sparse autoencoders (SAEs) have produced important insights in language model interpretability, but their utility on transformers trained on scientific data remains underexplored. We extend the SAE-plus-causal-intervention toolkit to an epigenomics foundation model, EpiBERT, and ask whether it internally encodes a biologically meaningful contrast: in vitro (cell line) vs. in vivo (primary tissue) chromatin context. We train layer-wise Sparse Autoencoders (SAEs) with BatchTopK activations across six matched ATAC-seq conditions spanning blood, liver, and lymph lineages, introduce the *Context Divergence Score* (CDS)—a contrastive t-statistic applicable to any probed transformer—to identify context-specific features, and validate them through causal ablation, linear context-steering, and three-level biological annotation (ChromHMM, HOMER, GO:BP). We find a depth-stratified context representation: context-specific features grow 3.8-fold from early to late layer (57 $\rightarrow$ 215 Bonferroni-significant), mirroring the late-layer concentration of high-level features in language model SAEs. Causal ablation of CDS-selected features yields a large effect, context-steering closes 11.2% of the prediction gap at $4.5\times$ above random, and biological annotation grounds the discovered features in lineage-defining transcription factors. These results demonstrate that the SAE methodology transfers cleanly from language to genomics, and that CDS provides a general primitive for identifying contrastive concepts in any probed transformer. Code is available at `https://github.com/nicoleching515/gene_expression_predictions`

[1]California Institute of Technology, Pasadena, California. Correspondence to: Nicole Ching <nching@caltech.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. *Copyright 2026 by the author(s)*

1. Introduction

The past decade has produced a wave of genomics foundation models trained on DNA sequence, epigenomic signal, or both (Zhou & Troyanskaya, 2015; Avsec et al., 2021; Ji et al., 2021; Nguyen et al., 2024; Dalla-Torre et al., 2023). These models achieve state-of-the-art performance on regulatory element prediction, variant effect scoring, and gene expression forecasting. Yet a fundamental tension underlies their deployment: training data comes predominantly from immortalised cell lines, while the settings of greatest clinical interest involve primary tissues with physiological chromatin states that differ substantially from those of cell lines. Cell lines exhibit aberrant epigenomes shaped by transformation, long-term culture, and clonal selection. ENCODE has generated extensive ATAC-seq data for both cell lines (*e.g.*, K562, HepG2, GM12878) and matched primary tissues (*e.g.*, hematopoietic stem cells, liver tissue, naive B cells), providing a natural paired axis of biological variation. If a chromatin model's internal representations conflate these two regimes, predictions trained on cell-line data may systematically fail to generalise—a concern with direct consequences for variant interpretation, regulatory annotation, and drug target identification.

This question matters beyond a particular biological application. SAE-based interpretability has rapidly become a standard methodology for analyzing transformer internals, with most work to date focused on language models. Whether the empirical regularities observed there—monosemantic feature dictionaries, depth-stratified concept hierarchies, causally manipulable representation subspaces—transfer to transformers trained on fundamentally different data modalities is an open question with broad implications for the generality of the SAE framework. Genomics foundation models, with their well-characterized inputs (DNA, ATAC-seq) and biologically grounded ground-truth concepts (cell lineages, regulatory elements, transcription factor identity), offer a particularly tractable test bed.

Mechanistic interpretability (Elhage et al., 2021; Olah et al., 2020) offers tools to locate specific internal features responsible for encoding given concepts. Specifically, sparse autoencoders (SAE) (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024) decomposes dense

transformer activations into an over-complete set of approximately monosemantic features, enabling fine-grained analysis of model computations. This framework has produced important insights in language model interpretability but has never been applied to an epigenomics foundation model.

This work. We apply SAEs with BatchTopK activations (Bussmann et al., 2024) to EPIBERT, probing three layer depths across six matched ATAC-seq conditions spanning three tissue lineages. We introduce the CDS, a contrastive t-statistic, to score context-specific features, validate them causally via targeted ablation and context-steering, and provide three-level biological annotation via ChromHMM state mapping, HOMER (Heinz et al., 2010) TF motif enrichment, and GO:BP enrichment (rGREAT; McLean et al. 2010). Together these provide the first mechanistic and biological characterisation of context encoding in an epigenomics foundation model.

Contributions.

1. We extend SAE-based mechanistic interpretability beyond language models to an epigenomics foundation model, demonstrating that the standard methodology (BatchTopK SAEs, causal ablation, linear feature steering) transfers cleanly to a new data modality.
2. The CDS metric: a statistically principled, contrastive feature scoring method applicable to any probed transformer, with explicit Bonferroni-corrected significance criteria and aggregate cross-condition scoring.
3. Strong causal evidence (Cohen's d=1.79, p=2.98×10^{-8}) that CDS-selected features mediate prediction differences, with linear context-steering closing 11.2% of the prediction gap as a working representation-engineering intervention.
4. Three-level biological annotation (ChromHMM, HOMER, GO:BP) confirms discovered features capture lineage-defining biology (HNF4A/FOXA2 in liver, SPI1/RUNX1 in blood, EBF1/PAX5 in lymph), providing ground-truth validation of SAE feature interpretability in a non-language domain.
5. Depth stratification: context-specific feature count grows 3.8-fold from early to late layer, mirroring the late-layer concentration of high-level features in language model SAEs and providing cross-domain evidence for this regularity.

2. Related Work

Epigenomics foundation models. Enformer (Avsec et al., 2021) extended Basenji to model long-range chromatin interactions over 200 kb windows using multi-head self-attention. Hyena-DNA (Nguyen et al., 2024) and Nucleotide Transformer (Dalla-Torre et al., 2023) scale to full-genome resolution. These models are evaluated predominantly on held-out cell types or genetic variants, with limited attention to in vivo generalisation or to the internal representations that mediate it.

Mechanistic interpretability and SAEs. The superposition hypothesis (Elhage et al., 2022) proposes neural networks pack more features than dimensions by exploiting approximate sparsity. SAEs operationalise this via sparse, near-monosemantic feature dictionaries (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). BatchTopK (Bussmann et al., 2024) stabilises SAE training by enforcing a fixed per-sample L_0 without auxiliary losses. Recent work has begun extending SAEs beyond language models to vision transformers and other scientific data modalities, examining whether the empirical regularities observed in LM SAEs (monosemanticity, depth-stratified concepts, causal steerability) generalize. Causal methods—activation patching and representation engineering—have validated features in language models (Meng et al., 2022), and our work extends this validation regime to a transformer trained on chromatin accessibility signal.

Interpretability in regulatory genomics. Saliency maps and in silico mutagenesis (Linder et al., 2022), activation maximisation (Novakovsky et al., 2023), and linear probing (Tenney et al., 2019) identify input-space features but do not characterise the internal feature space of the model. Our work operates at the sparse feature dictionary level, enabling causal interventions and multi-modal biological annotation inaccessible to input-attribution approaches.

3. Background

3.1. EPIBERT

EPIBERT is a BERT-style transformer encoder taking a 131,072 bp genomic window tiled into non-overlapping bins. Each bin is represented by a one-hot DNA sequence embedding projected to a d-dimensional vector, augmented by a scalar ATAC-seq signal (library-size normalised read depth). The model is pre-trained to predict per-bin chromatin accessibility across ENCODE cell types, with hidden dimension d=1024 at all layers. We denote total depth as L and analyse the pre-trained checkpoint at training step 45.

3.2. Sparse Autoencoders with BatchTopK

An SAE reconstructs $\boldsymbol{h}\in\mathbb{R}^d$ via:

$$\hat{\boldsymbol{h}} = W_{\text{dec}}\,\sigma_k(W_{\text{enc}}\boldsymbol{h} + \boldsymbol{b}_{\text{enc}}) + \boldsymbol{b}_{\text{dec}}, \tag{1}$$

where $W_{\text{enc}} \in \mathbb{R}^{m \times d}$, $W_{\text{dec}} \in \mathbb{R}^{d \times m}$ with unit-normalised columns, and σ_k retains the top-k activations per sample and zeros the rest. Training minimises $\mathcal{L} = \mathbb{E}[\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|_2^2]$. We use $m = 8192$, $d = 1024$ ($8\times$ expansion), $k = 64$, giving $L_0 = 64$ exactly by construction.

3.3. Context Divergence Score

We introduce the Context Divergence Score (CDS) to principally distinguish between in-vivo enriched and in-vitro enriched features.

Definition 3.1 (CDS)**.** *For* SAE *feature j and context pair $c = (\text{vitro}, \text{vivo})$, the* CDS *is the two-sample t-statistic over n evaluation windows:*

$$\text{CDS}_j^c = \frac{\bar{a}_j^{\text{vivo}} - \bar{a}_j^{\text{vitro}}}{\sqrt{s_j^{2,\text{vivo}}/n + s_j^{2,\text{vitro}}/n}}. \tag{2}$$

Feature j is vivo-enriched *if* $\text{CDS}_j^c > 0$ *and passes Bonferroni correction ($\alpha = 0.05$) over all m features;* vitro-enriched *if* $\text{CDS}_j^c < 0$*. Aggregate* CDS *is the mean absolute t-statistic over all three pairs.*

4. Methods

4.1. Data Pipeline

We downloaded ATAC-seq BAM files (hg38) from ENCODE (ENCODE Project Consortium, 2012) for six biosamples forming three matched pairs (Table 1).

Table 1. Matched biosample pairs. All data from ENCODE; hg38.

Pair	In vitro (cell line)	In vivo (tissue)
Blood	K562	Hematopoietic Stem Cell
Liver	HepG2	Liver Tissue
Lymph	GM12878	Naive B Cell

We sampled 200 non-overlapping 131,072 bp windows from chromosomes 8–9 (seed 42) as a held-out evaluation set, normalised signals per-sample (library-size then quantile normalisation to a pooled reference), and binned to EPIBERT's input resolution. Peak Jaccard overlaps between matched in vitro/in vivo pairs: 0.093 (blood), 0.327 (liver), 0.227 (lymph), confirming substantial divergence in raw accessibility landscapes that motivates a representation-level analysis.

4.2. Activation Collection and SAE Training

We ran 1,200 forward passes (200 windows $\times$ 6 conditions), extracting per-window mean-pooled hidden states at three depths. For SAE training, 10,000 windows per condition were sampled via random chromosomal sampling, giving 60,000 pooled vectors per layer. Three SAEs were trained in total—one per layer (early, mid, late), each pooled across all six conditions—for 50,000 steps (batch 4,096; lr 3×10^{-4}; 1,000-step warmup; resampling every 2,500 steps) on an NVIDIA H100 80 GB GPU. Final normalised MSE: 0.0013 / 0.0027 / 0.0073 (early/mid/late), all below the 0.05 threshold, with $L_0 = 64$ exactly for all layers. Approximately 23.5% of features ($\approx$1,925 per SAE) are stably active; dead-feature cycling between $\sim$23% and $\sim$75% is expected for BatchTopK at $8\times$ expansion and is not a failure mode. Full hyperparameters are in Appendix A.

4.3. Causal Ablation

We zeroed the top-k vivo-enriched features (by aggregate CDS) from the mid-layer hidden state during in vivo forward passes, re-ran the remaining transformer layers, and measured prediction shift $\Delta\hat{y} = \|\hat{y}_{\text{ablated}} - \hat{y}_{\text{orig}}\|_1$ against random ablation (5 seeds) across $k \in \{5, 10, 25, 50, 100\}$. We tested the global effect ($n = 25$ pairs) with the Wilcoxon signed-rank test and Cohen's d.

4.4. Context Steering

We applied a linear intervention to mid-layer activations during in vitro forward passes:

$$\boldsymbol{h}' = \boldsymbol{h} + \alpha \sum_{j \in \mathcal{V}} f_j(\boldsymbol{h}) \boldsymbol{d}_j - \beta \sum_{j \in \mathcal{C}} f_j(\boldsymbol{h}) \boldsymbol{d}_j, \tag{3}$$

where $\mathcal{V}$ and $\mathcal{C}$ denote vivo- and vitro-enriched feature sets, $f_j(\boldsymbol{h})$ is the SAE activation of feature j, and $\boldsymbol{d}_j$ is its decoder direction. Gap closure $GC = (\hat{y}^{\text{steered}} - \hat{y}^{\text{vitro}})/(\hat{y}^{\text{vivo}} - \hat{y}^{\text{vitro}})$ measures the fraction of the in vitro/in vivo prediction gap closed by the intervention, with ceiling $GC = 1$ at direct ATAC context swap. Sweep: $\alpha \in \{1.5, 2.0, 3.0, 5.0\}$, $\beta \in \{0.0, 0.25, 0.5\}$ vs. a random-direction baseline.

4.5. Biological Annotation

BED files of top-activation windows for the top-50 CDS features per layer per condition were annotated with three complementary tools. **ChromHMM**: intersected against 15-state Roadmap Epigenomics segmentations for K562, HepG2, and GM12878 (hg38 lift-over) via `bedtools intersect`; per-condition state proportions and net active (TssA, TssAFlnk, Enh, EnhG, Tx, TxFlnk) vs. repressed (ReprPC, ReprPCWk, Het) fractions shown in Figures 5 and 6. **HOMER** (Heinz et al., 2010): `findMotifsGenome.pl` (200 bp windows, hg38, repeat-masked, matched genomic background) for enriched TF binding motifs; results in Figure 7. **GO:BP** (McLean et al., 2010): rGREAT on the same BED files; results in Figure 8.

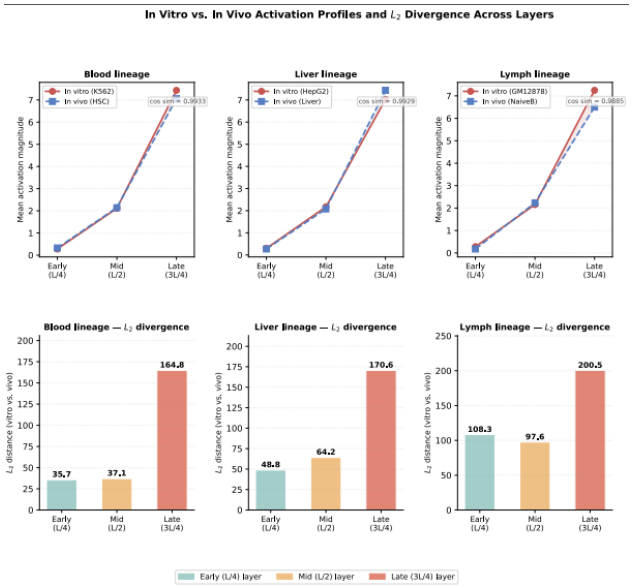

Figure 1. **In vitro vs. in vivo activation profiles and L_2 divergence (Fig. 1).** *Top:* mean activation magnitude per condition (red = in vitro; blue = in vivo); late-layer cosine similarity ($\geq$0.988) annotated. *Bottom:* L_2 distance between matched mean vectors. Blood/liver divergence is late-layer dominated; lymph divergence is high at all depths, reflecting EBV transformation of GM12878.

5. Results

5.1. Raw Activation Divergence Motivates SAE Analysis

Figure 1 shows mean activation magnitude (top) and L_2 divergence between matched mean vectors (bottom) across layers. All conditions track closely at early and mid depths but diverge sharply at the late layer. For blood and liver, divergence surges at late (164.8 and 170.6), a 4.4–4.5$\times$ amplification. The lymph pair diverges earlier: L_2=108.3 at early, remaining elevated at mid (97.6), consistent with EBV immortalisation of GM12878 driving broad epigenomic remodelling at every depth. Despite large L_2 distances, late-layer cosine similarities remain $\geq$0.988—representations share a common orientation while diverging along specific sparse dimensions, precisely the regime where SAE decomposition is most informative.

5.2. Context-Specific SAE Features Increase with Depth

Table 2 reports Bonferroni-corrected context-divergent feature counts. The signal grows 3.8-fold from 57 (0.70%) at early to 215 (2.62%) at late, with CDS std growing 3.4$\times$ (0.195 $\rightarrow$ 0.657) and maximum aggregate |CDS| rising from 8.87 to 24.11. This depth-stratified pattern mirrors findings in language model SAEs, where higher-level features concentrate in later layers (Templeton et al., 2024).

Figure 2 disaggregates by pair and direction. Vitro-enriched divergence is dominated by the liver pair at the late layer

Table 2. Bonferroni-significant context-divergent features by layer. Feature count grows 3.8-fold; CDS std grows 3.4$\times$.

LAYER	SIG.	%	CDS STD	MAX |CDS|
EARLY ($L/4$)	57	0.70	0.195	8.87
MID ($L/2$)	82	1.00	0.229	8.30
LATE ($3L/4$)	**215**	**2.62**	**0.657**	**24.11**

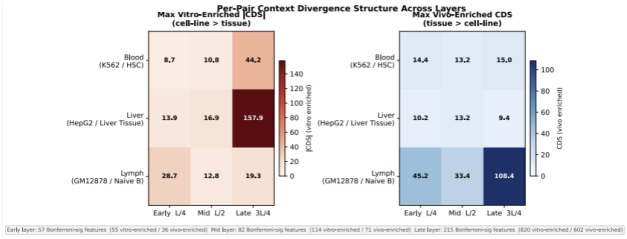

Figure 2. **Per-pair CDS divergence structure across layers (Fig. 2).** Each cell: max |CDS| per tissue pair (row) and layer depth (column). *Left (red):* max vitro-enriched |CDS|; liver/late dominates at 157.9, reflecting HepG2 carcinoma chromatin deregulation. *Right (blue):* max vivo-enriched CDS; lymph dominates at early (45.2) and late (108.4), reflecting EBV transformation of GM12878. Bottom: Bonferroni-significant counts per layer.

(|CDS|=157.9, $>3\times$ blood or lymph), reflecting HepG2's carcinoma chromatin deregulation. Vivo-enriched divergence shows the lymph pair with high enrichment at both early (45.2) and late (108.4), consistent with EBV-driven epigenomic distance. Only 0.1–1.1% of features show the same direction across all three pairs simultaneously, confirming features are predominantly lineage-specific rather than universal vitro/vivo markers. Vitro-enriched features outnumber vivo-enriched at every layer (early: 55 vs. 36; mid: 114 vs. 71; late: 820 vs. 602), a $\sim$1.5$\times$ asymmetry discussed in Section 6.

5.3. Targeted Ablation Confirms Causal Relevance

Figure 3 shows the ablation dose-response at the mid layer. Fold-enrichment peaks at k=25 (Table 3): targeted ablation $\Delta\hat{y}$=12.36 vs. random 4.53$\pm$0.24, a 2.73$\times$ fold. The global test is highly significant: Wilcoxon p=2.98$\times10^{-8}$ (BH-FDR p=1.19$\times10^{-7}$), Cohen's d=1.79. Top-activation ablation outperforms targeted in raw shift (up to 2.80$\times$ at k=10), confirming CDS isolates contextual specificity rather than general salience. The fold-enrichment plateau beyond k=25 suggests a compact set of $\sim$25 features carries the bulk of the causal signal at this layer.

5.4. Context Steering Partially Closes the Prediction Gap

The best steering setting (α=2.0, β=0.5) achieves median *GC*=0.112 (95% CI $[0.000, 0.171]$), 4.5$\times$ above random (*GC*=$-$0.025; Figure 4, Table 4). At $\alpha\geq$3.0, *GC* collapses to zero while windows with *GC*>0.5 rises to $\sim$31%, in-

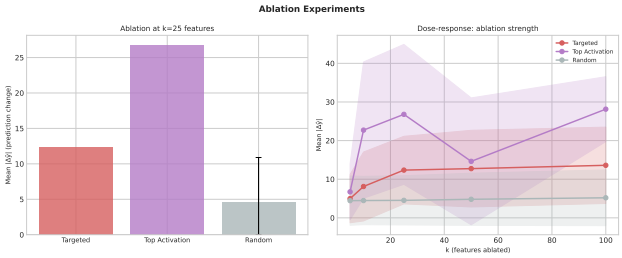

Figure 3. **Ablation dose-response at the mid layer (Fig. 3).** Targeted ablation of vivo-enriched features (blue) vs. random (grey, mean ± std, 5 seeds) and top-activation (orange dashed). Peak fold at k=25 (2.73×). Global Wilcoxon test (n=25): p=2.98×10^{-8}, Cohen's d=1.79.

Table 3. Ablation dose-response (mid layer). Best fold at k=25 in bold.

k	TARGETED $\Delta\hat{y}$	RANDOM $\Delta\hat{y}$	FOLD
5	4.97	4.45±0.13	1.12×
10	8.11	4.48±0.08	1.81×
25	**12.36**	**4.53±0.24**	**2.73×**
50	12.74	4.80±0.29	2.65×
100	13.60	5.19±0.36	2.62×

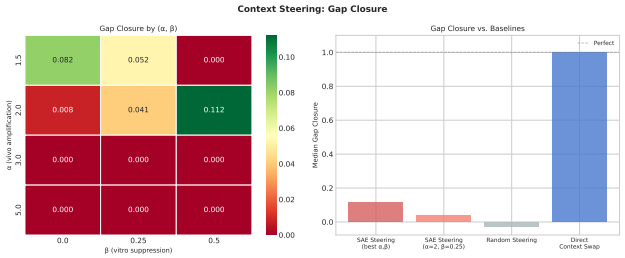

Figure 4. **Context-steering gap closure at the mid layer (Fig. 4).** Median *GC* across α and β. Best (α=2.0, β=0.5; blue): *GC*=0.112, 4.5× above the random baseline (grey, −0.025). At $\alpha\geq$3.0, *GC* collapses as the steered state leaves the training distribution.

Table 4. Steering results at selected settings (mid layer, n=200). Full sweep: Appendix C.

α	β	MED. *GC*	95% CI	FRAC>0.5
1.5	0.00	0.082	[0.00, 0.25]	29.0%
2.0	0.25	0.041	[0.00, 0.14]	28.5%
2.0	**0.50**	**0.112**	[0.00, 0.17]	**29.5%**
3.0	ANY	0.000	≈0	∼31%
RANDOM	—	−0.025	—	—
DIRECT SWAP	—	1.000	[1.0, 1.0]	100%

dicating out-of-distribution behaviour. Random steering consistently yields negative *GC* (mean −0.027), confirming the positive signal at $\alpha\leq$2.0 is not a random artefact. Notably, ∼29% of windows reach *GC*>0.5 at the best setting, identifying loci where context-specific regulatory logic is largely additive in feature space and therefore amenable to lightweight in silico tissue calibration.

5.5. ChromHMM Annotation: Active Enhancers vs. Polycomb Repression

State composition (Figure 5). Across all lineages and layer depths, in vivo features consistently show higher proportions of active regulatory states—particularly **Enh** (enhancers) and **TssAFlnk** (flanking active TSS)—relative to matched in vitro features. The contrast is sharpest for the liver pair at the late layer: Liver Tissue features are dominated by enhancer states, while HepG2 features show elevated **ReprPC** (Polycomb repressed) and **Het** (heterochromatin), directly reflecting HepG2's PRC2-mediated silencing of hepatocyte-specific enhancers. The lymph pair shows elevated **TxWk** (weak transcription) in Naive B cell features even at the early layer, consistent with broader transcriptional activation in mature B cells vs. EBV-immortalised GM12878.

Enrichment heatmap and active/repressed balance (Figure 6). The enrichment heatmap confirms **Enh** is the dominant discriminating state across late-layer vivo conditions, while **Quies** and **ReprPC** are elevated in vitro. The active vs. repressed balance shows in vivo features trend toward net active chromatin at the late layer; late-layer vitro liver features are the most repressed, matching the maximum |CDS|=157.9.

5.6. HOMER Motif Enrichment: Lineage-Defining Transcription Factors

Figure 7 shows the HOMER known-motif enrichment heatmap ($-\log_{10}(p)$) for top-50 context-divergent features per condition. The pattern is strikingly lineage-specific and aligns with established regulatory landscapes for these cell types.

Liver pair. Vivo-enriched (Liver Tissue) features are most significantly enriched for **HNF4A**, **FOXA2**, and **HNF1A**—master regulators of hepatocyte identity—with enrichment concentrated at the late layer, consistent with the depth-stratified CDS signal. Vitro-enriched (HepG2) features show enrichment for **AP-1** (FOS/JUN), **SP1**, and **E2F** family motifs, characteristic of the proliferative programme of hepatocellular carcinoma. This motif-level contrast directly explains the ChromHMM finding: HNF4A and FOXA2 are pioneer factors that open hepatocyte-specific enhancers, which HepG2 has silenced via Polycomb.

Blood pair. Vivo-enriched (HSC) features are enriched for **SPI1/PU.1**, **RUNX1**, and **C/EBP**α—master regulators of multi-potent hematopoietic progenitor identity. Vitro-enriched (K562) features show enrichment for **GATA1**,

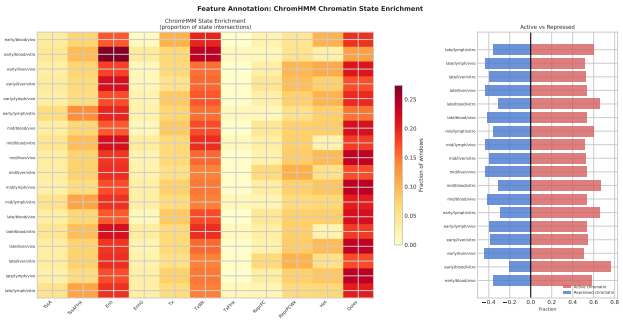

Figure 5. **ChromHMM state composition of top context-divergent SAE features (Fig. 5).** Stacked bars: state intersection proportions (in vivo vs. in vitro) per tissue pair (columns) and layer depth (rows). In vivo: higher Enh, EnhG, TssAFlnk. In vitro: elevated ReprPC, Het—sharpest for liver at the late layer, consistent with HepG2 Polycomb repression of hepatocyte enhancers.

GATA2, and **TAL1/SCL**, reflecting K562's erythroid and megakaryocyte lineage bias.

Lymph pair. Vivo-enriched (Naive B Cell) features are enriched for **EBF1**, **PAX5**, and **E2A/TCF3**—the core TFs required for B cell commitment. Vitro-enriched (GM12878) features show enrichment for **IRF4**, **NF-κB/RELA**, and **AP-1** motifs, consistent with the EBV-driven transformation programme activating NF-κB signalling. The layer-spanning enrichment of EBF1/PAX5 motifs in vivo features (significant at both early and late layers) is consistent with the high L_2 divergence of the lymph pair at all depths.

5.7. GO:BP Enrichment: Tissue-Specific Biological Processes

Figure 8 shows the GO:BP enrichment dotplot, providing pathway-level confirmation of the biological identities encoded by context-specific features.

Liver. Vivo-enriched features are significantly enriched for *regulation of lipid metabolic process*, *bile acid biosynthesis*, *fatty acid beta-oxidation*, and *hepatocyte differentiation*—the core metabolic functions of mature hepatocytes, reflecting the HNF4A/FOXA2 regulatory programme. Vitro-enriched features are enriched for *mitotic cell cycle*, *DNA replication*, and *regulation of cell proliferation*, consistent with HepG2's carcinoma state. This GO-level contrast is the strongest in the dataset and explains both the maximum |CDS| and the extreme ChromHMM contrast for the liver pair.

Blood. Vivo-enriched features are enriched for *hematopoietic progenitor cell differentiation*, *regulation of myeloid cell differentiation*, and *stem cell maintenance*, consistent with the SPI1/RUNX1 programme of HSCs. Vitro-enriched features are enriched for *regulation of cell cycle* and *telomere organisation*, reflecting K562's CML-driven proliferative programme.

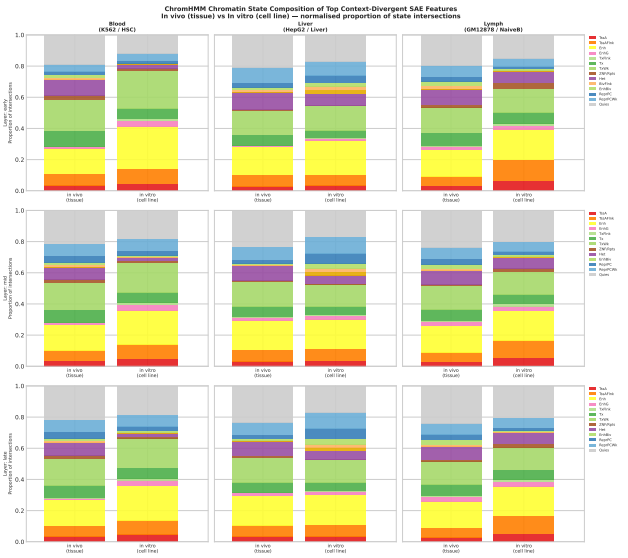

Figure 6. **ChromHMM enrichment heatmap and active/repressed balance (Fig. 6).** *Left:* state enrichment across 18 conditions (3× layers × 3× pairs × 2 contexts). Enh/TssAFlnk prominent in late-layer vivo; Quies/ReprPC elevated in late-layer vitro liver. *Right:* net active (red) vs. repressed (blue) balance. Late-layer vitro liver features are most repressed, matching the maximum |CDS| in the dataset.

Lymph. Vivo-enriched (Naive B Cell) features are enriched for *B cell receptor signalling pathway*, *adaptive immune response*, and *lymphocyte differentiation*, consistent with EBF1/PAX5 motif enrichment. Vitro-enriched (GM12878) features are enriched for *response to virus*, *NF-κB signalling*, and *regulation of lymphocyte apoptosis*. Notably, the enrichment of viral response processes in GM12878 features confirms that the lymph pair's layer-spanning L_2 divergence is driven by fundamental cell identity differences introduced by EBV immortalisation, not culture conditions.

6. Discussion

Context representations are real, sparse, and fully biologically validated. Our results establish that EPIBERT encodes the in vitro/in vivo distinction in a small, identifiable set of SAE features concentrated in later network layers. The three-level biological annotation closes the interpretability loop: ChromHMM confirms features fire over tissue-specific active regulatory elements; HOMER identifies the specific TF motifs that define each lineage's regulatory identity; and GO:BP confirms the features correspond to tissue-specific biological processes. Together, these three independent annotation layers demonstrate that EPIBERT has learned a genuine representation of in vivo chromatin biology, not merely statistical correlates of cell identity.

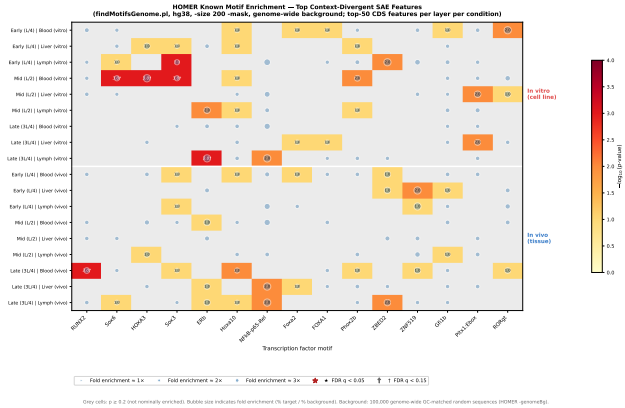

Figure 7. **HOMER TF motif enrichment of top context-divergent SAE features (Fig. 7).** Heatmap of $-\log_{10}(p)$ across all 18 conditions (rows: 3 layers $\times$ 3 pairs $\times$ 2 contexts; columns: top enriched motifs). In vivo: lineage-defining TFs (HNF4A/FOXA2, SPI1/RUNX1, EBF1/PAX5). In vitro: proliferative and transformation TFs (AP-1/E2F, GATA1/TAL1, IRF4/NF-κB). Grey = not significant (p>0.05 Bonferroni).

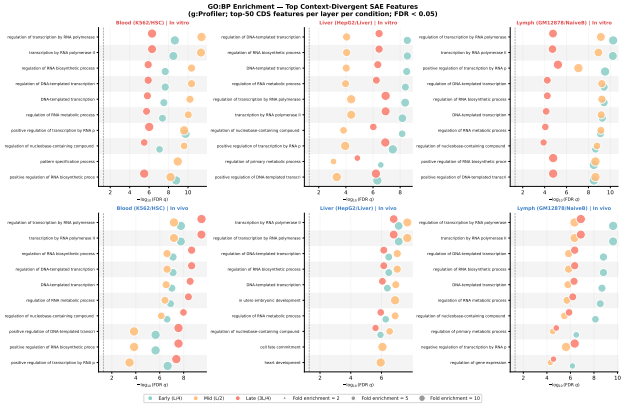

Figure 8. **GO:BP enrichment of top context-divergent SAE features (Fig. 8).** Dotplot faceted by tissue pair (columns) and context (rows). Dot size = fold enrichment; colour = layer (teal: early, orange: mid, red: late); x-axis = $-\log_{10}$(FDR q). Dashed line: q=0.05 threshold. In vivo: tissue-specific biological processes (hepatocyte metabolism, HSC differentiation, B cell activation). In vitro: proliferative and transformation-associated processes (cell cycle, viral response, NF-κB signalling).

The in vitro/in vivo distinction is mechanistically explained per lineage. The three-level annotation reveals distinct biological mechanisms per lineage. For liver, the model encodes the loss of hepatocyte-specific enhancers (HNF4A/FOXA2-driven, enriched for lipid metabolism and hepatocyte differentiation) through HepG2's Polycomb-mediated silencing and acquisition of a proliferative carcinoma programme (AP-1/E2F, cell cycle). For blood, the contrast is between multi-potent HSC identity (SPI1/RUNX1, hematopoietic differentiation) and K562's erythroid/megakaryocyte lineage bias (GATA1/TAL1). For lymph, the contrast is between primary B cell identity (EBF1/PAX5, B cell receptor signalling, adaptive immunity) and the EBV-transformed lymphoblastoid state (IRF4/NF-κB, viral response). Critically, the GO enrichment for viral response processes in GM12878 vitro features confirms that the lymph pair's early-layer L_2 divergence is driven by a fundamental cell identity difference, not a superficial accessibility change. The lineage-specificity itself is a load-bearing finding for downstream applications: a single "in vitro vs. in vivo" correction cannot exist, because the mechanistic axis differs per tissue.

Causal evidence is robust and biologically grounded. The ablation result (Cohen's d=1.79, p=2.98$\times10^{-8}$) establishes that identified features causally mediate prediction differences. The HOMER and GO annotations now reveal *what* these causal features are computing: windows enriched for HNF4A, SPI1, and EBF1 binding motifs and for tissue-specific metabolic/differentiation processes. This is a mechanistically stronger claim than correlation-based interpretability.

Steering gap closure has a biologically interpretable limit. The 11.2% gap closure reflects a minimal intervention on $\sim$185 features at a single mid-layer. The biological annotation suggests why the remaining gap persists: the cell-line/tissue difference involves not only the presence of enhancer-associated features but their precise genomic co-localisation with lineage-defining TF binding sites (HNF4A at liver-specific enhancers, SPI1 at hematopoietic regulatory elements) that require correct genomic context to function. Multi-layer iterative steering and tissue-specific fine-tuning are promising directions for closing the remaining gap.

Vitro enrichment asymmetry. Vitro-enriched features outnumber vivo-enriched by $\sim$1.5$\times$ at all layers. The GO:BP enrichment for cell cycle and proliferation processes in vitro features confirms this reflects a genuine biological asymmetry: cell lines maintain a proliferative regulatory programme absent in resting primary tissues, and ENCODE contains more cell-line ATAC-seq experiments than tissue experiments. Both factors contribute to the asymmetry and are independently testable predictions for future pre-training data design.

Implications for tissue-specific genomics and transfer learning. Two actionable interventions follow from our results. First, fine-tuning on primary tissue ATAC-seq with frozen early layers should efficiently update late-layer enhancer features and TF motif representations while preserving regulatory grammar learned from the cell-line corpus. The late-layer dominance of the contrast for blood and liver suggests that relatively few fine-tuning steps may be required. Second, CDS-informed activation steering

could serve as a lightweight in silico tissue calibration—particularly for the $\sim$29% of loci where steering achieves $GC>0.5$, which may correspond to regions where tissue-specific regulatory logic is primarily additive.

Implications for SAE methodology beyond language models. Our results contribute three observations to the question of whether SAE methodology generalizes beyond language models. First, the empirical depth-stratification of high-level features—a regularity established for LM SAEs (Templeton et al., 2024)—reproduces cleanly in a transformer trained on a fundamentally different modality (DNA + chromatin accessibility), suggesting it may be a general property of SAE-decomposed transformer representations rather than an artifact of language data. Second, the strong causal effect of CDS-selected features (Cohen's d=1.79) and the partial steerability of context representations (11.2% gap closure) confirm that the standard interventional toolkit—ablation, linear feature steering, decoder-direction injection—works on a chromatin foundation model with no methodological modification. Third, the availability of well-characterized biological ground truth (TF motifs, ChromHMM states, GO terms) makes epigenomics models a uniquely tractable test bed for validating SAE feature interpretability claims, complementing the unsupervised feature discovery typical in LM settings.

Limitations. The 200-window evaluation set is underpowered for rare or weakly divergent features; BH-FDR analysis on a larger window set would reveal more. Reference genome inputs used synthetic random sequences and JASPAR motif scores were set to zero; both cancel in the CDS difference. The blood pair showed unexpectedly low peak Jaccard (0.093), possibly due to the small HSC BAM (1.1 GB vs. $\sim$10 GB for K562), which may slightly underpower blood-specific CDS estimates. HOMER and GO analyses used pooled feature BED files; per-feature annotation to identify the most specific regulators for individual top-CDS features is reserved for future work. Our findings are based on a single epigenomics foundation model (EPIBERT) at one pre-trained checkpoint; whether the depth-stratified, lineage-specific context structure we observe generalizes to Enformer, HyenaDNA, or Nucleotide Transformer is an important question for follow-up work.

7. Conclusion

We have demonstrated that EPIBERT encodes in vitro vs. in vivo chromatin context in a sparse, depth-stratified, causally relevant, and biologically interpretable set of SAE features. Context-specific feature count grows 3.8-fold from early to late layer (57 $\rightarrow$ 215); causal ablation yields a large effect (Cohen's d=1.79, p=2.98×10^{-8}); and context-steering closes 11.2% of the prediction gap at $4.5\times$ lift over random. Three-level biological annotation reveals the mechanisms behind these findings: an HNF4A/FOXA2 enhancer vs. Polycomb contrast in liver; SPI1/RUNX1 progenitor vs. GATA1/TAL1 erythroid in blood; and EBF1/PAX5 B cell identity vs. IRF4/NF-κB EBV transformation in lymph. These results establish SAE-based mechanistic interpretability as a productive framework for genomics foundation model analysis, contribute cross-domain evidence for the generality of SAE empirical regularities, and provide CDS as a general-purpose primitive for identifying contrastive concepts in any probed transformer.

Acknowledgments

We thank Michelle Lu for helping shape the early framing of the paper.

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18:1196–1203, 2021.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bussmann, B., Leask, P., and Nanda, N. BatchTopK SAEs: Achieving stable sparsity without auxiliary losses. *arXiv preprint arXiv:2412.06410*, 2024.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Gressent, N. L., Carrier, A. H., Trop, M., de Almeida, B. P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework

for transformer circuits. *Transformer Circuits Thread*, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489: 57–74, 2012.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38 (4):576–589, 2010.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Linder, J., Bogard, N., Rosenberg, A. B., and Seelig, G. Interpreting neural networks for biological sequences by learning stochastic masks. *Nature Machine Intelligence*, 4:41–54, 2022.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28:495–501, 2010.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., Ermon, S., Re, C., and Baccus, S. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Novakovsky, G., Dexter, N., Libbrecht, M. W., Bhatt, D., and Wasserman, W. W. Obtaining genetics insights from deep learning via explainability. *Nature Reviews Genetics*, 24:125–137, 2023.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015.

A. SAE Training Details

Table 5. Full SAE training configuration and QC metrics.

Hyperparameter	Value
d_{input}	1024
d_{latent}	8192 (8× expansion)
k (BatchTopK)	64
Training steps	50,000
Batch size	4,096
Learning rate	3×10^{-4}
Warmup steps	1,000
Resample interval	2,500 steps
Hardware	NVIDIA H100 80 GB
Wall time per SAE	≈19 min
Final QC	
Norm-MSE (early/mid/late)	0.0013 / 0.0027 / 0.0073
L_0 (all layers)	64.0 (exact)
Stably active features	≈23.5% per layer

Dead-feature fraction cycles between ∼23% (post-resample) and ∼75% (pre-resample). This is expected for BatchTopK at 8× expansion where each sample activates only 0.78% of features. Reconstruction quality is excellent (norm-MSE <0.01) and $L_0{=}k$ exactly; the cycling is not a failure mode.

B. Cross-Layer Feature Overlap

Figure 9. **Cross-layer Jaccard overlap of top-50 CDS features (Fig. A1).** Early ∩ mid: 1 shared feature (Jaccard 0.020, $p{=}0.037$, BH-FDR $q{=}0.074$, not significant after correction); early ∩ late and mid ∩ late: 0 shared features. Near-zero overlap across all layer pairs indicates that context-specific features at each depth are largely non-overlapping; however, given only 50 features are selected from 8,192, this result is close to the random expectation and should be interpreted cautiously.

Table 6. Cross-layer Jaccard overlap of top-50 CDS features. Neither overlap survives BH-FDR correction.

LAYER PAIR	JACCARD	p (HYPERGEOM.)	BH q
EARLY ∩ MID	0.020	0.037	0.074
EARLY ∩ LATE	0.000	1.000	1.000
MID ∩ LATE	0.000	1.000	1.000

C. Full Steering Sweep

Table 7. Complete (α, β) sweep (mid layer, $n{=}200$). Bold = best.

α	β	Med. *GC*	95% CI	Frac>0.5
1.5	0.00	0.082	[0.000, 0.247]	29.0%
1.5	0.25	0.052	[0.000, 0.173]	26.5%
1.5	0.50	0.000	[−0.000, 0.026]	18.5%
2.0	0.00	0.008	[0.000, 0.107]	29.5%
2.0	0.25	0.041	[0.000, 0.138]	28.5%
2.0	**0.50**	**0.112**	[0.000, 0.171]	29.5%
3.0	0.00	0.000	[−0.000, 0.001]	31.5%
3.0	0.25	0.000	[−0.000, 0.002]	31.0%
3.0	0.50	0.000	[−0.000, 0.005]	30.5%
5.0	0.00	0.000	[−0.000, 0.012]	31.0%
5.0	0.25	0.000	[−0.000, 0.009]	31.0%
5.0	0.50	0.000	[−0.000, 0.009]	30.5%
Random	—	−0.025	—	—
Direct swap	—	1.000	[1.000, 1.000]	100%

D. Data Pipeline Details

ATAC-seq signals were obtained via `pysam` pileup at 128 bp resolution, normalised by sequencing depth (reads per million), and quantile-normalised across all six conditions. Reference genome inputs used synthetic random one-hot sequences; JASPAR motif scores were set to zero. Both constants are identical across all six conditions and cancel in all CDS computations.